# 面向微博话题的"主题+观点"词条抽取算法研究*

姚兆旭　马　静
(南京航空航天大学经济与管理学院　南京 211106)

**摘要**: 【目的】自动抽取微博话题信息, 从主题及观点两个维度整合揭示微博话题内容与观点。【方法】将主题模型应用于微博话题中, 结合改进的 TF-IDF 算法, 构建主题特征词向量; 基于特征词向量中特征词之间的相关度, 自动抽取主题词汇链; 引入情感词典, 抽取主题观点, 无监督构建"主题+观点"词条。【结果】使用爬虫工具抽取 2014 年 6 月–2015 年 6 月期间 4 个特定热门微博话题事件的微博共 24 598 条, 抽取"主题+观点"词条, 平均准确率达到 80.3%, 召回率为 76.7%。【局限】数据量依旧较小, 主题模型对于微博短文本的特征抽取效果仍需提高。【结论】本文算法可以准确且有效地描述话题事件内容及其相应观点。

**关键词**: 文本挖掘　词条抽取　主题模型　微博话题
**分类号**: TP391　G350



## 1 引　言

随着互联网的普及与发展, 博客、微博和社交网络等网络平台成为网民获取信息的重要来源。截至 2015 年 6 月, 中国网民规模达 6.68 亿[1], 微博用户规模为 2.49 亿, 其中有 64.6%的用户参与过热门话题讨论[2]。由此可见, 微博社区已成为重要的舆情传播平台, 其中微博话题已成为用户针对话题事件获取信息、表达观点的重要渠道。但是由于微博话题中信息鱼龙混杂以及微博自身短文本、结构松散的特性, 因此迫切需要一种合适的组织模式或框架, 帮助用户在信息过载时迅速抽取与表达微博话题信息, 多维度展现舆情内容。

话题信息的抽取与表达可以追溯到话题识别与跟踪(Topic Detection and Tracking, TDT)的话题检测阶段, 话题检测的主要任务[3]是检测和组织话题, 通常用于应对信息过载问题。近年来国内外关于话题信息抽取方法的研究主要从数据挖掘方法与 NLP 文本挖掘方法两方面切入。数据挖掘方法主要从结构化、半结构化数据中抽取信息。Becker 等[4]利用 Twitter 一段时间内的历史数据, 通过聚类算法获得事件簇, 提取事件簇特征, 并利用支持向量机模型在线识别新文本。Popescu 等[5]针对某一类特定产品的评论信息, 通过计算评论中的名词与该类产品表征词间的点互信息(PMI), 使用贝叶斯分类来提取产品特征。NLP 文本挖掘方法从非结构化的开放文本中发现新知识, 并将其转换为可理解的有用信息。Rttier 等[6]针对 Twitter 自身特性, 提出开放领域事件抽取方法, 利用潜在变分模型发现重要的事件类别。然而, 由于微博话题文本相较于传统文本内容简短、结构松散、数据稀疏性严重, 因此传统的文本抽取方法并不适用于微博话题文本。

随着主题模型[7]的提出, 为了解决上述方法存在的问题, 越来越多的学者将LDA模型引入到话题抽取

通讯作者: 姚兆旭, ORCID: 0000-0003-0906-1670, E-mail: 269708110@qq.com。

*本文系国家自然科学基金项目"基于演化本体的网络舆情自适应跟踪方法研究"(项目编号: 71373123)、江苏高校哲学社会科学研究重点项目"基于超网络的江苏教育微博舆情多元意见演化模型及应用研究"(项目编号: 2015ZDIXM007)和南京航空航天大学基本科研业务费重大项目培育基金项目"基于'模型–数据双驱动'的复杂社会网络行为大数据分析方法研究"(项目编号: NP201630X)的研究成果之一。

与表达的研究中。LDA 模型是三层贝叶斯分布的概率模型，将话题中隐含主题信息通过特征词概率分布来表示。为进一步提升模型适用性与话题抽取效果，有学者在传统 LDA 模型基础上引入情感因素[8]、话题热度[9]、作者信息[10]、微博间用户关系[11]等外部因素，进行微博短文本研究，并取得良好效果。目前基于主题模型的话题信息抽取与表达主要从话题标签抽取[12]、话题线索化[13]、话题演化[14]等方面进行研究。寇宛秋等[12]提出一种基于种子词的话题标签抽取方法，对话题特征词权重重排序，抽取种子词，采用 Bootstrapping 思想，生成关键短语集合，最后泛化选择话题标签，表述话题内容。Ramage 等[15]针对 Twitter 中博文的内容特征，利用标签 LDA(Labeled LDA)模型将博文内容映射到 4 个维度，抽取标签，反映话题信息。Darling 等[16]提出 PoSLDA 模型，在 LDA 模型和 HMMLDA 模型的基础上进一步扩展，将文档中的词汇分为三个类别(形容词、动词和名词)表示话题涉及的事物、动作和描述信息。闫泽华[17]在 LDA 模型基础上，调整单词权重，考虑背景词与 N 元短语的因素，抽取新闻线索标签。这些研究都是从话题内容方面抽取话题信息，并没有考虑引入观点维度改善微博话题信息抽取与表达，更加全面展现话题信息。

为了进一步提升话题信息抽取与表达效果，本文设计面向微博话题的“主题+观点”表达模型，并提出一种无监督的“主题+观点”词条抽取算法。实验结果表明本文算法在不同微博话题中均取得较好的效果，从多维度反映微博话题中各主题信息及主题观点。

## 2 面向微博话题的“主题+观点”词条模型

微博话题的微博语义信息可以分为两类：话题事件的客观描述信息与主观观点信息。本文综合微博话题自身特性，提出面向微博话题的“主题+观点”词条的话题表达模型，“主题+观点”词条由主题词汇链与主题观点两部分组成，主题词汇链以词汇链的形式表征微博话题中各主题事件内容信息，主题观点反映用户对主题事件的观点倾向。

定义 1 主题词汇链 $\text{Lexicalchain}\{k_n\}$ 由一组具有代表性的单词或者短语组成，根据特征词集合中词汇的相关性 $\text{cor}(w_i, w_j)$ 自动构建生成，用以表征微博话题中主题事件的内容信息。

定义 2 主题观点 $\text{Viewpoint}\{j_n\}$ 是代表主题 $z_i$ 观点信息的观点词，反映网民对主题事件的观点意见。

定义 3 “主题+观点”词条 Entry(n) 表示从主题内容与主题观点两个维度揭示话题信息，由主题词汇链 $\text{Lexicalchain}\{k_n\}$ 与主题观点 $\text{Viewpoint}\{j_n\}$ 构成，模型结构如下所示:

$$\text{Entry}(n) = \text{Lexicalchain}\{k_n\} + \text{Viewpoint}\{j_n\} \quad n = 1, 2, \cdots, K \quad (1)$$

其中，主题词汇链 $\text{Lexicalchain}\{k_n\}$ 表示第 n 个主题 $z_n$ 的主题词汇链，主题观点 $\text{Viewpoint}\{j_n\}$ 表示对应主题信息的观点信息，K 表示主题数目。

## 3 无监督的“主题+观点”词条抽取算法

当前针对话题信息抽取的普遍解决思路是有监督、半监督或无监督的文本挖掘方法。有监督方法仅具理论价值，因为实际应用中难以拟出合适的训练集构建分类器。本体作为一种有效的形式语义模型和知识表示形式，近年来在话题抽取方面也有一定应用，但构建话题相关本体往往采用半监督方式，需要引入大量领域信息，准确度不高，多为原型系统，未能走向应用[18]。然而无监督探测算法则兼具较少先验需求与较强泛化能力，更符合话题抽取的实际情境。

本文提出一种无监督的微博话题信息抽取算法，主要分为以下三步:

(1) 根据主题特征词在话题中不同主题间代表度的差异，调整特征词权重，构建主题特征词向量;

(2) 在特征词向量的基础上，依照特征词之间相关度，无监督生成主题词汇链表征主题内容信息;

(3) 引入情感词典，构建观点词集合，结合步骤(2)中主题词汇链与观点词的观点强度，自动抽取主题观点，用以描摹主题事件观点倾向。最终主题词汇链与主题观点构成“主题+观点”词条，将微博话题信息从文本维度降维表示，从主题事件内容与观点的维度描述话题信息。

### 3.1 基于改进 TF-IDF 算法的主题特征词向量构建

在 LDA 主题模型中，通过降维将话题信息从海量文本空间变换到主题空间，将一组词汇的概率分布表示话题中一个主题(Topic)，即通过一组特征词描述话题中一个主题事件。假设话题文本集合 $D = \{d_1, d_2, \cdots d_n\}$，其中 $V = \{w_1, w_2, \cdots w_n\}$ 为词汇集

合，潜在主题事件集合为$Z=\{z_1,z_2,\cdots z_n\}$，主题建模后得到主题–词概率分布θ与文档主题概率分布φ，其中$p(w_j|z_k)$表示在主题$z_k$下词汇$w_j$对主题的贡献度，即$w_j$属于主题$z_k$的概率。

LDA模型假设每个词汇权重相同，但实际上每个词汇在各个主题中代表度并不相同。传统词汇代表度的计算通常使用TF-IDF算法，但TF-IDF存在无法有效识别高频关键词与无法筛选均匀分布的关键词的问题，本文借鉴文献[12]的思想，在传统TF-IDF算法基础上引入覆盖度与特征度，使主题特征词与背景词区分出来。

覆盖度$coverage_{i,j}$表示词汇在文档集合上的覆盖程度，覆盖度高的词语在语料中更具有代表性，覆盖度用包含词语的全部文档数$N_i$除以总文档数N来表示，计算公式如下：

$$coverage_{i,j}=\frac{N_i}{N} \tag{2}$$

特征度$characteristic_i$反映词汇所在文本在某个主题中代表程度。$p(z_i|d_n)$为主题–文本概率分布，代表词汇$w_i$所在的文本属于主题$z_i$的概率，为包含词汇$w_i$的微博$d_n$代表主题$z_i$的概率，计算公式如下：

$$characteristic_i=\frac{\sum_{i=1}^{n}p(z_i|d_n)}{N} \tag{3}$$

结合改进TF-IDF算法，其表达式如下：

$$weight_{i,j}=p(w_i|z_j)\times\log\frac{|W_{num}|}{|W_{num}^i|}\times coverage_{i,j}\times characteristic_i \tag{4}$$

其中，$W_{num}$表示文档词汇总数，$W_{num}^i$表示$w_i$的词频数量。可见，词汇在文档中出现次数越多，包含词汇文档数目越少，代表度越高，同一主题中覆盖度越大，特征度越高的特征词更能代表主题语义。本文通过改进的TF-IDF算法计算词汇权重后，按权重$weight_{i,j}$数值从大到小排序，选取前n个特征词，组成主题特征词向量，则调整后主题$z_n$的特征词向量表示如下：

$$\overline{z_n}=\{(w_1,weight_{1,j}),(w_2,weight_{2,j}),\cdots,(w_n,weight_{n,j})\} \quad w_1\cdots w_n\in V \tag{5}$$

基础LDA模型与权重计算结果对比如图1所示，上方为LDA主题建模结果，下方为权重计算后的主题特征向量。




图1　基础LDA模型与权重计算结果对比

### 3.2　基于特征词向量的主题词汇链生成

词语常常围绕特定主题描述话题信息，这些围绕某个主题，在语义上相互联系的词语集合，称为词汇链。本文用特征词之间相关度的大小反映不同词汇间语义关联的强弱。常用的词汇间相关度计算公式如下：

$$cor(w_i,w_j)=\frac{c(w_i,w_j)}{c(w_i)+c(w_j)-c(w_i,w_j)}-\frac{c(w_i)\times c(w_j)}{N\times N} \tag{6}$$

其中，$c(w_i,w_j)$为$w_i$与$w_j$在同一窗体中出现的频率，$c(w_i)$，$c(w_j)$为各自的词频，N为全部文档数目。由于在计算中可能得到词汇为负相关，考虑到$c(w_i),c(w_j)$相比于N通常较小，因此公式(6)中后半部分可以忽略。传统相关度计算中忽视了$w_i$与$w_j$自身权重的影响，当权重较高的特征词间相关度高时，其组成短语更易于反映主题信息。由此本文相关度计算公式改进为如公式(7)所示，在原相关度计算公式基础上引入特征词权重$weight_{i,j}$。

$$cor(w_i,w_j)=\sum weight_{i,j}\left(\frac{c(w_i,w_j)}{c(w_i)+c(w_j)-c(w_i,w_j)}\right) \tag{7}$$

其中，$c(w_i,w_j)$在本文中取词汇在同一微博中共现次数，$c(w_i)$，$c(w_j)$为$w_i$与$w_j$在语料文本中出现

次数。可见，若相关性为正值，则说明两词相关，正值越大，则相关度越高。当两个权重高的特征词的共现概率高时，则词汇间相关度变大，其组成短语更能准确反映话题语义信息。

文献[19]认为，新闻领域通常使用新闻六要素来描述一个事件，即：内容(What)、人物(Who)、地点(Where)、时间(When)、原因(Why)及如何(How)。文献[20]认为评论中的观点持有者一般是由命名实体，提出借助于命名实体识别技术来获取观点持有者。本文借鉴上述思想，认为描述话题事件的文本通常包含名词词性的词汇，由此选择特征词集合中的名词词性的特征词 $w_i^n$ 作为种子词，用以自动生成主题词汇链。主题词汇链 $Lexicalchain\{k_i\}$ 依据种子词 $w_i^n$ 与特征词向量中其他特征词的相关度和特征词的权重等因素生成。当种子词 $w_i^n$ 与特征词的相关度 $cor(w_i^n, w_j)$ 大于阈值后，将种子词与特征词组成短语 $P^t$ 加入词汇链候选集合 $P_i$ 中，短语的权重更新为词汇的权重之和。迭代计算后，从中选取权重最大的短语 $P^t$ 作为主题词汇链，即 $Lexicalchain\{k_i\} = \arg\max_{0<k<K} P^t(weight_{i,j} \mid z = i)$，主题词汇链生成算法如下：

```
Input: 特征词集合 V_i，特征词短语集合 P_i
Output: 主题词汇链 Lexicalchain{k_i}
①Set P_i=V_i;
②For each w_j in V_i
    Calculate weight_i,j
    Add all w_i^n in list;
  End for;
③For each w_i^n in list
    For each P^t in P_i
      Calculate cor(w_i^n, w_j)
      If cor(w_i^n, w_j) ≥ 阈值
        Set (w_i^n, w_j) as a phrase into P_i
    End for
  End for
④For each P^t in P_i
    Set maximum weight P^t as Lexicalchain{k_i}
  End for
```

### 3.3　基于情感词典的主题观点抽取

观点抽取指利用计算机技术自动分析网络中带有观点信息的句子或文档，从中提取出用户所表达的观点或态度。话题文本中的观点倾向主要通过观点词传递，观点词多为情感词，其中观点词体现为观点倾向(褒义、贬义和中立)与观点强度两个维度。

本文借鉴大连理工情感词本体[21]结果，构建情感词典，大连理工情感词本体通过三元组来描述，具体如下：

$$Lexicon = (B, R, E) \tag{8}$$

其中，B 表示词汇基本信息，R 表示词汇之间同义关系，E 代表词汇情感信息，分别从情感分类、极性、强度三个维度描述。通过候选情感词与基准情感词在大规模语料中点互信息(PMI)判定情感强度，强度分为 1, 3, 5, 7, 9 这 5 个等级。在微博文本中，越来越多的用户使用微博表情代替文字信息，表达个人观点，在本文的实验语料中，含有微博表情的文本占 46.7%。由此在大连理工情感词本体的基础上，对情感词典进行扩充，加入微博常用表情。将微博表情以[表情内容]的形式表示，如“[鼓掌]”“[爱你]”，存入情感词典，以其文本内容代表表情语义。经过处理，情感词典共有单词共 28 466 个，褒义词 16 074 个，贬义词 12 392 个，情感强度参考情感词本体，也分为 1, 3, 5, 7, 9 这 5 个等级。

参照情感词典，假设在主题特征词向量中标记出 m 个观点词，当前主题观点词向量表示为 $SW = \{(sw_1, sweight_{sw_1}), (sw_2, sweight_{sw_2}) \cdots (sw_m, sweight_{sw_m})\}$，$sw_m$ 表示观点词，$sweight_{sw_m}$ 为对应观点强度。然而主题观点的表达不仅与观点强度相关还与观点次和主题内容的紧密程度有关。本文采用主题词汇链 $Lexicalchain\{k_i\}$ 表示主题事件的语义信息。因此，将观点抽取过程转化为观点词与主题词汇链的相关度的计算过程。本文定义主题观点的观点值 $Q_i$，其计算公式如下：

$$Q_i = sweight_{sw_i} \times \sum_{i}^{n} cor(sw_i, w_j) \mid sw_i \in SW, w_j \in Lexicalchain\{k_i\} \tag{9}$$

可见，当观点词观点强度越大，同时该观点词与主题内容相关度越高，则观点词更能代表主题观点。对所有被标注出的观点词 sw，满足以下两种情况之一，则自动抽取为主题观点 $View\{j_n\}$：

(1) 对于 $\forall sw_i \in Lexicalchain\{k_i\}$，$Viewpoint\{j_n\} = \arg\max\{weight_{sw_i}\}$，$i \in \{1, 2, \cdots n\}$；

(2) 如果 $\exists sw_i \in SW$ 且 $sw_i \notin Lexicalchain\{k_i\}$，使得 $Viewpoint\{j_n\} = argmax\{Q_i\}$，$i \in \{1, 2, \cdots n\}$。

即，在条件(1)中，如果主题词汇链中存在情感词，选取权重最大的特征词作为主题观点；在条件(2)中，选取在观点词集合中观点值最大的观点词作为主题观点。

## 4 实验设计

### 4.1 实验设置

本文使用爬虫工具抽取 2014 年 6 月到 2015 年 6 月热门微博话题事件微博共 24 598 条，其中关于"成都女司机被打"话题微博共 9 230 条，关于"尼泊尔地震"微博共 6 932 条，关于"长江客轮沉没"话题共 4 367 条，"李娜产女"话题共 4 069 条。预处理阶段，使用中国科学院计算技术研究所汉语分词系统 NLPIR2015[①] 对微博正文文本分词并进行词性标注。根据哈工大停用词表去除停用词，同时去掉微博短链以及低频词，保留名词、动词、形容词作为候选词。

### 4.2 实验结果

实验设置参数 $\alpha = 50/K$，$\beta = 0.01$，吉布斯采样的迭代次数为 1 000 次，其中 K 为设置的话题数量。为了分析话题数量的设置对于 LDA 话题建模的影响，采用 Perplexity 指标对实验结果进行衡量。Perplexity 是度量话题模型性能的常用指标和衡量方法，表示预测数据时的不确定度，取值越小表示性能越好，计算公式如下：

$$Perplexity(W) = \exp\left\{-\frac{\sum_m \ln p(w_m)}{\sum_m N_m}\right\} \quad (10)$$

其中，W 为测试集，$w_m$ 是测试集中可观测到的词语，$N_m$ 为词语数。逐步递增话题数 K 进行实验，按照公式(10)计算不同话题值下 LDA 话题的混杂度。随着话题数逐渐增加，Perplexity 值不断降低，如图 2 所示：

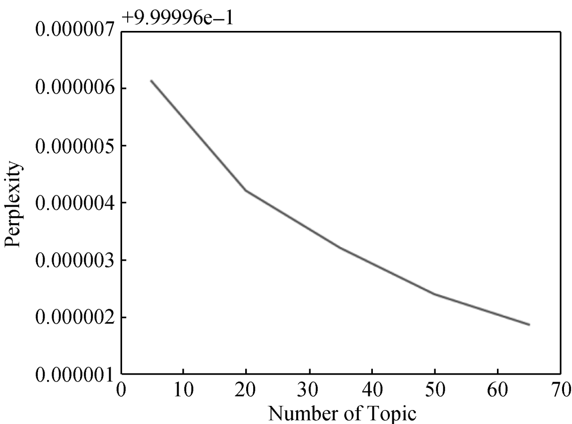



图 2 Perplexity 分布图

本文最终选取 K=50，实验结果如表 1 所示：

表 1 "主题+观点"词条抽取结果

| 事件名称 | 权重计算结果 | 主题词汇链 | 主题观点 | "主题+观点"词条 |
|---|---|---|---|---|
| 李娜产女 | 李娜 公布 顺利 英文名 小脚 Alisa 喜讯 孩子 发布 英文 | 李娜公布 | 喜讯 | 李娜公布+喜讯 |
| | 李娜 产女 祝福 人生 哈哈 祝 满贯 快乐 手 成长 | 李娜产女 | 祝福 | 李娜产女+祝福 |
| | 李娜 恭喜 成为 妈妈 中国 [鼓掌] 曝光 姜山 冠军 升级 | 李娜成为妈妈 | 恭喜 | 李娜成为妈妈+恭喜 |
| 成都司机被打 | 女司机 变道 成都 视频 记录 车 行车 殴打 遭 曝光 | 女司机变道 | 殴打 | 女司机变道+殴打 |
| | 女司机 惨遭 暴打 分析 骂 成都 [笑 cry] 男司机 评论 事 | 女司机暴打 | 暴打 | 女司机暴打+骂 |
| | 女司机 慈善 做 母亲 视频 机构 [笑 cry] 事 女儿 搞 | 女司机慈善 | [笑 cry] | 女司机慈善+[笑 cry] |
| 尼泊尔地震 | 中国 游客 回国 尼泊尔 地震 [心] 嫦 爱 想 愿 | 中国游客回国 | [心] | 中国游客回国+[心] |
| | 尼泊尔 地震 [祈祷] 中国 国际 [心] 汶川 加油 力量 [位置] | 尼泊尔地震 | [祈祷] | 尼泊尔地震+[祈祷] |
| | 西藏 受灾 自治区 捐赠 救灾 影响 受 波及 房屋 前往 | 西藏受灾 | 捐赠 | 西藏受灾+捐赠 |
| | 救援队 救援 搜救 队员 中国 国际 应急 帐篷 小时 废墟 | 救援队搜救 | 救援 | 救援队搜救+救援 |
| 长江客船沉没 | 长江 愿 平安 客船 沉没 乘客 入 公司 星 沉 | 长江客轮 | 平安 | 长江客船+平安 |
| | 长江 倾覆 客轮 安息 生命 加油 珍惜 生 长 爱 | 长江客轮 | 倾覆 | 长江客轮+倾覆 |

① http: //ictclas.nlpir.org/.

对于某一微博话题内容，往往含有多个不同主题内容即子话题。由实验结果可以看出 LDA 模型在主题挖掘领域中具有良好的效果，主题间具有较高独立性，主题特征词具有较高的概括性，充分反映出不同主题间的文本内容，有效去除垃圾微博对话题事件的影响。例如"尼泊尔地震"事件中，分别反映出"中国旅客回国"、"西藏地区受灾"、"救援队救援"等主题信息，将围绕微博话题的不同主题信息有效区分开来。

在特征词集合构建中，根据改进的 TF-IDF 算法提升话题特征词的权重，降低无关的背景词的权重，突出主题特征。例如，在"李娜产女"的话题中"中国"、"冠军"是与话题相关性较低词汇，经过计算，本文方法减少了无关词汇影响，同时提升了"李娜"、"女儿"等词汇的权重，更好地反映话题内容。

观点词抽取反映出用户对话题内容的观点意见，展示事件发展过程中对话题中不同主题的态度。例如在"成都女司机被打"事件中，事件开始女司机被打的观点为"骂"，表达对打人事件的谴责，经过事件发展，女司机借口去做慈善而违章变道，则该主题事件中的观点是"[笑 cry]"，表达反讽与不相信。

本文的"主题+观点"词条能够较好地反映话题信息，基本覆盖话题中各主题事件内容，从主题内容信息与主题观点两个维度表征话题。例如"长江沉船沉没"事件中，自动抽取词条"长江客船+平安"、"长江客轮+倾覆"，虽然主题词汇链相同，但属于微博话题中不同讨论的内容，前者为客船祈福，后者是话题事件的描述报道。

**4.3　对比实验**

对本文方法同新浪微博话题标签与文献[12]提出的方法进行比较。新浪微博话题标签一般由人工编辑，作为对微博话题事件的概述。文献[12]抽取每个主题的种子词，迭代产生关键短语集合，最后泛化选择话题标签，描述话题信息。结果如表 2 所示:

表 2　对比实验结果

| 事件名称 | 新浪微博话题标签 | 文献[12]话题标签 | "主题+观点"词条 |
|---|---|---|---|
| 尼泊尔地震 | 尼泊尔 8.1 级地震 | 尼泊尔地震<br>中国救援队<br>西藏地震 | 尼泊尔地震+[祈祷]<br>西藏受灾+捐赠<br>救援队搜救+救援<br>中国游客回国+[心] |
| 李娜产女 | 李娜产女 | 李娜公布孩子 | 李娜产女+祝福<br>李娜成为妈妈+恭喜 |
| 成都司机被打 | 成都女司机<br>变道遭殴打 | 女司机成都狂殴<br>女司机驾驶 | 女司机变道+殴打<br>女司机惨遭+暴打 |
| 长江客轮 | 长江客轮沉没<br>长江客轮倾覆 | 全国人民<br>长江客轮 | 长江客船+平安<br>长江客轮+倾覆 |

从对比结果可以看出本文方法能够准确抽取表达出话题内容及观点。例如"尼泊尔地震"话题中，由于话题事件爆发突然，事件讨论相对集中，用户观点基本一致。新浪微博简单表述为"尼泊尔 8.1 级地震"，缺少对话题中各主题事件的多维度表达，文献[12]的话题标签只是单纯描述出话题事件内容，如"尼泊尔地震"、"西藏地震"，而本文对话题语料抽取出"中国游客回国+[心]"、"西藏受灾+捐赠"等，在表述主题信息同时，反映用户相应主题观点。

与文献[12]话题标签抽取方法相比，本文在反映主题内容信息同时，也反映出事件相应观点倾向，以便于用户了解话题全貌。例如，尼泊尔地震事件中，虽然两种方法都抽取出"尼泊尔地震"，但本文方法在表述话题事件的同时也反映出提及地震事件的微博大部分是为地震灾区祈祷，展现用户表达的观点。同时在"尼泊尔地震"事件中，本文方法还抽取出在被困中国游客回国的事件，而文献[12]方法并没有挖掘出相关主题。

在部分话题语义表达中，本文方法不如新浪微博话题标签，例如新浪微博话题标签为"成都女司机变道遭殴打"，本文抽取的词条为"女司机变道+殴打"，与之相比信息完整性与语义通顺性都有所欠缺。同时，

由于用户在发表的微博中，越来越多使用表情图标代替文本以表达观点，但包含微博表情的微博往往无明显句法结构，因此包含微博表情的词条在对话题信息解释性方面受到影响。

本文采用准确率P、召回率R和F1对比文献[12]与本文方法抽取的效果，计算公式如下：

$$P = \frac{C_{correct}}{C_{extract}} \tag{11}$$

$$R = \frac{C_{correct}}{C_{standard}} \tag{12}$$

$$F1 = \frac{2PR}{P+R} \tag{13}$$

其中，$C_{correct}$为正确抽取结果数目，$C_{extract}$为所有抽取结果，而$C_{standard}$为所有人工标注词条数目，结果如表3所示：

表3 实验精度评测结果

| 事件名称 | 准确率P | | 召回率R | | F1 | |
|---|---|---|---|---|---|---|
| | 本文方法 | 话题标签 | 本文方法 | 话题标签 | 本文方法 | 话题标签 |
| 李娜产女 | 78.7% | 74.6% | 73.3% | 69.6% | 75.9% | 72.1% |
| 成都司机被打 | 75.3% | 68.7% | 71.7% | 66.2% | 73.4% | 67.4% |
| 尼泊尔地震 | 85.1% | 84.1% | 78.3% | 76.4% | 81.6% | 80.1% |
| 长江客轮沉没 | 80.3% | 78.9% | 76.7% | 77.8% | 78.4% | 78.3% |

从表3中可以看出，本文方法精度高于文献[12]方法，在"成都司机被打"的事件中，话题讨论多为网民自发参与，事件持续时间较长，伴随着事件的演化与发展，用户在不同阶段情感发生转变，文献[12]话题标签由于缺少观点维度表达，仅描述话题内容，因此准确率、召回率低于本文方法。在"李娜产女"的微博中存在大量"大满贯""中国网球"等无关背景词，在一定程度上干扰话题信息，本文方法有效降低了背景词对事件抽取的影响，因此准确率更高。同时在"尼泊尔地震"与"长江客轮沉没"的事件中由于话题中微博多为新闻报道类微博，具有通用格式，微博内容具有较高语义相似性，因此两种方法精度相近。

本文与文献[12]方法均采用LDA模型主题建模，但模型建模结果中存在部分主题语义表达不明确和掺杂大量垃圾微博信息的问题，如大量简单动词如"走"、"吃"、"去"、"爱"在同一个话题结果中，并不具备表达话题语义的能力，影响反映话题事件主要信息。

## 5 总结与展望

本文提出一种面向微博话题的"主题+观点"词条模型及其无监督抽取算法，该算法采用LDA建模，对词汇权重计算后，构建特征词集合；根据特征词间相关性，自动抽取出主题词汇链，表述主题内容信息；引入情感词典，得到主题观点，将主题词汇链与主题观点构建成"主题+观点"词条，从内容与观点两个维度表征微博话题信息。最后通过实验数据验证"主题+观点"词条在话题信息抽取与表达方面的实用性以及其无监督抽取算法的有效性。后续工作包括进一步准确抽取话题观点以及改进主题模型，提升主题抽取效果。

## 参考文献：


[1] 中国互联网络信息中心. 第36次中国互联网络发展状况统计报告[R/OL]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201507/P020150723549500667087.pdf. (China Internet Network Information Center. The 36th Statistical Report on the Network Development of China Internet [R/OL]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201507/P020150723549500667087.pdf.)

[2] 艾瑞咨询. 2014年中国微博用户行为研究报告[R/OL]. http://www.iresearch.com.cn/report/2183.html. (iResearch. The 2014 Research on China Weibo User Behavioral Report [R/OL]. http://www.iresearch.com.cn/report/2183.html.)

[3] 洪宇, 张宇, 刘挺, 等. 话题检测与跟踪的评测及研究综述[J]. 中文信息学报, 2007, 21(6): 71-87. (Hong Yu, Zhang Yu, Liu Ting, et al. Topic Detection and Tracking Review [J]. Journal of Chinese Information Processing, 2007, 21(6): 71-87.)

[4] Becker H, Naaman M, Gravano L. Beyond Trending Topics: Real-World Event Identification on Twitter[C]. In: Proceedings of the 5th International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain. AAAI Press, 2011.

[5] Popescu A M, Etzioni O. Extracting Product Features and Opinions from Reviews[A]. // Natural Language Processing and Text Mining[M]. Springer London, 2007.

[6] Ritter A, Mausam, Etzioni O, et al. Open Domain Event Extraction from Twitter[C]. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2012.

[7] Blei D M, Ng A Y, Jordan M I, et al. Latent Dirichlet

Allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[8] Lin C H, He Y L. Joint Sentiment/Topic Model for Sentiment Analysis [C]. In: Proceeding of the 18th ACM Conference on Information and Knowledge Management. New York: ACM, 2009: 375-384.

[9] 唐晓波，向坤. 基于 LDA 模型和微博热度的热点挖掘[J]. 图书情报工作，2014，58(5): 58-63. (Tang Xiaobo, Xiang Kun. Topic Mining Based on LDA Model and Popularity of Weibo[J]. Library and Information Service, 2014, 58(5): 58-63.)

[10] Rosen-Zvi M, Griffiths T, Steyvers M, et al. The Author-Topic Model for Authors and Documents [C]. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. 2012.

[11] 张晨逸，孙建伶，丁轶群. 基于 MB-LDA 模型的微博主题挖掘[J]. 计算机研究与发展，2011，48(10): 1795-1802. (Zhang Chenyi, Sun Jianling, Ding Yiqun. Topic Mining for Microblog Based on MB-LDA Model[J]. Journal of Computer Research and Development, 2011, 48(10): 1795-1802.)

[12] 寇宛秋，李芳. 基于种子词汇的话题标签抽取研究[J]. 中文信息学报，2013，27(5): 114-121. (Kou Wanqiu, Li Fang. Topic Label Extraction Based on Seed Word[J]. Journal of Chinese Information Processing, 2013, 27(5): 114-121. )

[13] 钱哲怡，李芳. 基于关键词和命名实体识别的新闻话题线索抽取[J]. 计算机应用与软件，2011，28(12): 168-171. (Qian Zheyi, Li Fang. Keyword and Name Entity Identification Based News Topic Thread Extraction[J]. Computer Applications and Software, 2011, 28(12): 168-171.)

[14] Hoffman M D, Blei D M, Bach F R. Online Learning for Latent Dirichlet Allocation[C]. In: Proceedings of the 24th Annual Conference on Neural Information Processing Systems. 2010.

[15] Ramage D, Hall D, Nallapati R, et al. Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora [C]. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore. 2009.

[16] Darling W, Song F．Probabilistic Topic and Syntax Modeling with Part-of-Speech LDA[OL]. arXiv: 1303.2826.

[17] 闫泽华. 基于 LDA 的新闻线索抽取研究[D]. 上海：上海交通大学，2012. (Yan Zehua. News Threading Based on LDA Model[D]. Shanghai: Shanghai Jiaotong University, 2012.)

[18] 王宇阳. 基于本体进化的自适应中文话题跟踪算法研究[D]. 南京：南京航空航天大学，2013. (Wang Yuyang. Research on Algorithm of Adaptive Chinese Topic Tracking Based on Ontology Evolution [D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2013.)

[19] 郭跚秀，吕学强，李卓. 基于突发词聚类的微博突发事件检测方法[J]. 计算机应用，2014，34(2): 486-490. (Guo Yixiu, Lv Xueqiang, Li Zhuo. Burstyn Topics Detection Approach on Chinese Microblog Based on Burst Words Clustering [J]. Journal of Computer Applications, 2014, 34(2): 486-490.)

[20] Kim S M, Hovy E. Determining the Sentiment of Opinions [C]. In: Proceedings of the 20th International Conference on Computational Linguistics. 2004.

[21] 陈建美. 中文情感词汇本体的构建及其应用[D]. 大连：大连理工大学，2008. (Chen Jianmei. The Construction and Application of Chinese Emotion Word Ontology [D]. Dalian: Dalian University of Technology, 2008.)

## 作者贡献声明:

姚兆旭：提出研究思路和研究方案，进行实验，论文撰写与修订；
马静：数据采集，扩展研究思路，论文审阅与修订。

## 利益冲突声明:

所有作者声明不存在利益冲突关系。

## 支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。

[1] 姚兆旭，马静. lda.zip. LDA 建模 JAVA 程序.
[2] 姚兆旭，马静. corpus.txt. 分词后的数据集.
[3] 姚兆旭，马静. ldaresult.towords. LDA 结果数据.
[4] 姚兆旭，马静. tfidfresult.xls. 特征词向量结果数据.
[5] 姚兆旭，马静. sentimentdictionary.sql. 情感词典.
[6] 姚兆旭，马静. finalresult.xls. 结果数据.

收稿日期: 2016-01-28
收修改稿日期: 2016-05-23

# Extracting Topic and Opinion from Microblog Posts with New Algorithm

Yao Zhaoxu    Ma Jing
(College of Economic and Management, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

**Abstract:** [**Objective**] This paper proposes an algorithm to extract topic and opinion information from the microblog posts automatically. [**Methods**] First, we used the improved TF-IDF algorithm to build the topic characteristic word vector. Second, we generated lexical chain for the topics based on the relevance among words of the vector. Finally, we extracted the topic and opinion information with the sentiment dictionary, and then generated the "topic+opinion" entries. [**Results**] We analyzed 24,598 Sina microblog posts of four trending events from June 2014 to June 2015 retrieved by a specially designed crawler. The precision and recall rates of the proposed method were 80.3% and 76.67%, respectively. [**Limitations**] The data size was small, the effect that the topic model extracted the feature about Weibo still required to be improved. [**Conclusions**] The proposed algorithm could effectively extract the "topic and opinion" information from micoblog posts.

**Keywords:** Text mining    Keyword extraction    Topic model    Microblog topic



## NISO 推出新项目，为图书馆电子内容创建灵活的 API 框架

国家信息标准组织(NISO)经过投票已经批准了一个新的项目，该项目的目标是增强现代化图书馆厂商技术的互操作性，以改善数字内容和电子书的访问。NISO 工作小组将基于 Queens Library 所开发的一套 API 需求，建立一个基础的供图书馆使用的 API 组。这个 API 组将实现用户和图书馆目前的需求，例如：更快的响应时间、灵活的发现和交付选项、更高的资源可利用性，以及电子资源和物理资源更加无缝的集成。

图书馆服务于读者时，应该给读者以优秀的用户体验，以及必要的便利性。NISO 的这一新项目试图将读者的图书馆体验和现代化工具，以及读者在生活中习惯使用的技术，特别是移动技术进行接轨。当今的图书馆使用了多种技术，其中的一些技术甚至依赖过时的、缓慢的通信协议来服务读者。通过建立 RESTful Web Services API 标准，以及移动扩展标准，图书馆行业将能够摒弃很多陈旧的、难以使用的工具，帮助图书馆在满足用户需求时能够有更强的灵活性。

志愿者工作组成员将以 NISO 推荐做法的形式提供一个基础框架，讨论图书馆如何提供和获取数据。这些图书馆相关的通讯和功能可能包括：定制的馆藏浏览、搜索和发现，用户认证，账户信息的传输，条形码管理，图书的借阅和归还，在线资源的流媒体化，以及其他利益相关者的需求。此外，工作组的工作还包括，创建一些使用推荐做法的服务案例，构建一个注册机制，帮助支持数据提供方和系统供应商沟通他们对基础框架的支持。完整的工作说明可在 NISO 官网(http://www.niso.org.)获取。

(编译自: http://www.niso.org/news/pr/view?item_key=e18a9742103bc945868a51a1e196e62b68879df6)

(本刊讯)